EMBO reports

Data sharing platforms and the academic evaluation system

Thijs Devriendt[1,*], Mahsa Shabani[2] & Pascal Borry[1]

Data sharing is both a prerequisite and an integral part of Open Science. To strengthen public access to research data and support open science goals, the NIH has issued a draft policy for data management and sharing (https://osp.od.nih.gov/wp-content/uploads/Draft_NIH_Policy_Data_Management_and_Sharing.pdf). Similar efforts have been made by the European Commission through, among others, funding various projects to develop sustainable infrastructures for data sharing. This includes collaborative projects such as euCanSHare and CINECA to create platforms for sharing data from disease and population cohorts across the EU and Canada.

Despite these efforts, significant challenges remain as many scientists are reluctant to broadly share their data. Experience with the WWARN data platform shows that active involvement and crediting of data contributors is crucial and that fears about getting scooped are a clear disincentive for sharing (http://www.ternyata.org/wp-content/uploads/2017/01/WWARNCaseStudy15Dec2016.pdf). Current reward and crediting mechanisms in academia are intensifying the challenges for data sharing, which has been noted by researchers and policy makers (Ali-Khan *et al*, 2017). Building and curating large-scale cohorts requires a lot of efforts and labor by physicians, data curators, data managers, and informaticians over many months or years. This work has at times been described as "invisible," as they are often not recognized in the academic reward system (Ankeny & Leonelli, 2015).

In response, the need for crediting data sharing has been put forward. Arguably, the traditional rewarding mechanisms including co-authorship for downstream data analysis may not seem fit for purpose (https://clarivate.com/webofsciencegroup/campaigns/global-research-report-multi-authorship-and-research-analysis/). Notably, systematically crediting all data generators has resulted in papers with hundreds of authors, contributing to the so-called "hyper-authorship" phenomenon and authorship inflation (Cronin, 2001). This trend has raised concerns over research integrity, such as the capacity of researchers to contradict prior conclusions of the data generators, how disputes over use of the methodology should be resolved, and the dilution of accountability (Bierer *et al*, 2017). Furthermore, there are concerns about the influence of hyper-authorship on popular metrics of scientific productivity (Hu *et al*, 2010).

We suggest an alternative approach, which leverages data-level metrics (DLMs) to capture and make data-sharing efforts visible. We conceptualize DLMs as indicators of scientific merit related to the production and (re-)use of datasets. For example, the number of downloads, metadata views, and data citations is already collected in many centralized repositories. The same mechanisms could be integrated into data-sharing platforms, although their distinct architecture and *modus operandi* for data sharing are not identical. Indicators of data quality, such as completeness or consistency, should also fall under DLMs. This is, in our view, particularly relevant to medical data although they could in principle be more broadly applied. The recording of these metrics can be integrated into emerging data-sharing platforms and eventually be used for academic evaluations. DLMs can thus be seen as complementary to recent proposals for specifying authorship in publications, such as introducing the Data Author designation (Bierer *et al*, 2017).

However, simply collecting DLMs through data-sharing platforms is insufficient as it does not embed the platform within the broader academic system. The platform should therefore systematically collect and transfer DLMs to digital spaces where they are visible for academic institutions and funding organizations. Without fulfilling these conditions, novel metrics will simply remain isolated in separate silos. Here, we make three recommendations on how connections between platform, funders, and academic institutions can be established to facilitate the use of DLMs.

First, ORCID profiles should display metrics related to datasets researchers have contributed to, so that these can be used in evaluating academic performance. Thus, datasets should be associated with a team of researchers or clinicians involved in data generation, curation, or other pre-analytical roles within the data-sharing platform. Scientific teams often collect cohort data over many years and the composition of the team might change over time. Therefore, the contributor roles attached to datasets need to be dynamic. If data are re-used, this should contribute to dataset metrics.

Second, infrastructures that support Open Access/Open Data such as OpenAIRE should receive metrics from data-sharing platforms and visualize DLMs for datasets over time. Notably, this option would fit well into the OpenAIRE Funder Dashboard that allows research funders and policy makers to monitor research outcomes. As such, this would provide funders with the possibility to see whether datasets have been uploaded and to observe indicators of the scientific

[1] Centre for Biomedical Ethics and Law, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium
[2] Metamedica, Faculty of Law and Criminology, University of Ghent, Ghent, Belgium
*Corresponding author. E-mail: thijs.devriendt@kuleuven.be

DOI 10.15252/embr.202050690 | EMBO Reports (2020) 21: e50690 | Published online 12 July 2020

productivity of all datasets derived from their funding. It would address the problem that many funders with Open Data policies do not actively follow-up on sharing, primarily owing to a lack of monitoring tools (https://zenodo.org/record/3401278#.XqlJ5cgzZPY). Furthermore, it would also make the enforcement of sharing mandates easier. Researchers can then be certain that sharing does not disadvantage them, as elevated DLM could increase their chances to acquire further funding (Sim *et al*, 2020).

Third, all collected data underlying DLMs should be made available for scientific research, so that they can be assessed, evaluated, and refined (Hicks *et al*, 2015). This is in line with the Open Science Policy Platform recommendation that: "[t]he data, metadata and methods that are relevant to research evaluation, including [...] citations, downloads and other potential indicators of academic re-use, should be publicly available for independent scrutiny and analysis by researchers, institutions, funders and other stakeholders"(https://ec.europa.eu/research/openscience/pdf/integrated_advice_opspp_recommendations.pdf). Thus, DLMs and information inherent to datasets such as cohort size, types of available data, phenotype richness, and study type should be made accessible. One way to realize this would be to pass on these data to the DataCite/Crossref Data Event service that is already collecting and collating similar metrics for datasets deposited within centralized repositories.

DLMs offer novel opportunities to incentivize data sharing, but they have their limitations. For instance, the use of data metrics may raise concerns about manipulation of metrics (https://ec.europa.eu/research/openscience/pdf/report.pdf). In the case of data-sharing platforms, data generators could for instance request access to their own data from several accounts to artificially increase the number of access requests. In addition, several technical and governance issues also need to be addressed: If data re-use takes place over several data-sharing platforms or central repositories, should these DLMs then be aggregated? Is it possible and desirable to attribute less credit for partial re-use of the dataset? Should sharing alone without re-use be in some way rewarded? These questions need to be discussed in view of the anticipated, downstream uses of DLMs in research evaluation.

Furthermore, the transition toward Open Science is a cultural change that involves the development of new policies, strategies, and the evaluation of outputs and work against open criteria. To successfully realize these changes, an environment of trust, collaboration, and commitment is required (Ayris *et al*, 2018). Notably, inertia against such changes can be expected owing to general conservatism in reward systems in academia, at times fueled by academics willing to preserve the system from which they have benefited previously (https://rio.jrc.ec.europa.eu/sites/default/files/report/MLE-OS-Report-3%20.pdf). Community engagement with researchers, funders, and institutions is necessary to raise support for the use of DLMs. All stakeholders involved should understand their uses, shortcomings, and limitations and be committed to their development and fair use.

The European Commission's Expert Group on Altmetrics recommended the development of alternative credit systems in support of Open Science and that greater investment should be made into the field of "meta-science"(https://ec.europa.eu/research/openscience/pdf/report.pdf). Notably, they call for "next-generation research data infrastructure[s], which can ensure greater efficiency and interoperability of data collection, and its intelligent and responsible use to inform research strategy, assessment, funding prioritization and evaluation in support of open science." In our view, data-sharing platforms are examples of such next-generation infrastructures and they could, in principle, be designed to advise research strategy and priorities. Moreover, many funders are open to other evaluation models for research. In the 2019 Scholarly Publishing and Academic Resources Coalition (SPARC) Report, approximately half of the funders have expressed support for or have signed the DORA Declaration, which calls for the abandonment of the Journal Impact Factor and to "consider the value and impact of all research outputs (including datasets and software) in addition to research publications [for the purposes of research assessment]"(https://zenodo.org/record/3401278#.XqlJ5cgzZPY).

Finally, active collaboration and dialogue between researchers, metrics developers, (bio)informaticians, and policy makers will be necessary to successfully tackle the incentive problems for Open Data. In addressing these issues, the onus should be on how data-sharing platforms can inform Open Data policies in the coming years. By influencing and shaping policies at an earlier stage, it can be ensured that scientists do contribute their data and receive proper credit for doing so. Data-sharing platforms are then rightfully recognized as indispensable components that can catalyze future data sharing and re-use in biomedical sciences.

Acknowledgements

This publication is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825903 (https://ec.europa.eu/programmes/horizon2020/en).

References

Ali-Khan SE, Harris LW, Gold ER (2017) *Elife* 6: e29319

Ankeny RA, Leonelli S (2015) Valuing data in postgenomic biology: how data donation and curation practices challenge the scientific publication system. In *Postgenomics: perspectives on biology after the genome*, Richardson SS, Stevens H (eds), pp 126 – 149. Durham, NC: Duke University Press

Ayris P, López de San Román A, Maes K, Labastida I (2018) *Leag Eur Res Univ* 24: 13

Bierer BE, Crosas M, Pierce HH (2017) *N Engl J Med* 376: 1684 – 1687

Cronin B (2001) *J Am Soc Inf Sci Technol* 52: 558 – 569

Hicks D, Wouters P, Waltman L, De Rijcke S, Rafols I (2015) *Nature* 520: 429 – 431

Hu X, Rousseau R, Chen J (2010) *J Inf Sci* 36: 73 – 85

Sim I, Stebbins M, Bierer BE, Butte AJ, Drazen J, Dzau V, Hernandez AF, Krumholz HM, Lo B, Munos B *et al* (2020) *Science* 367: 1308 – 1309